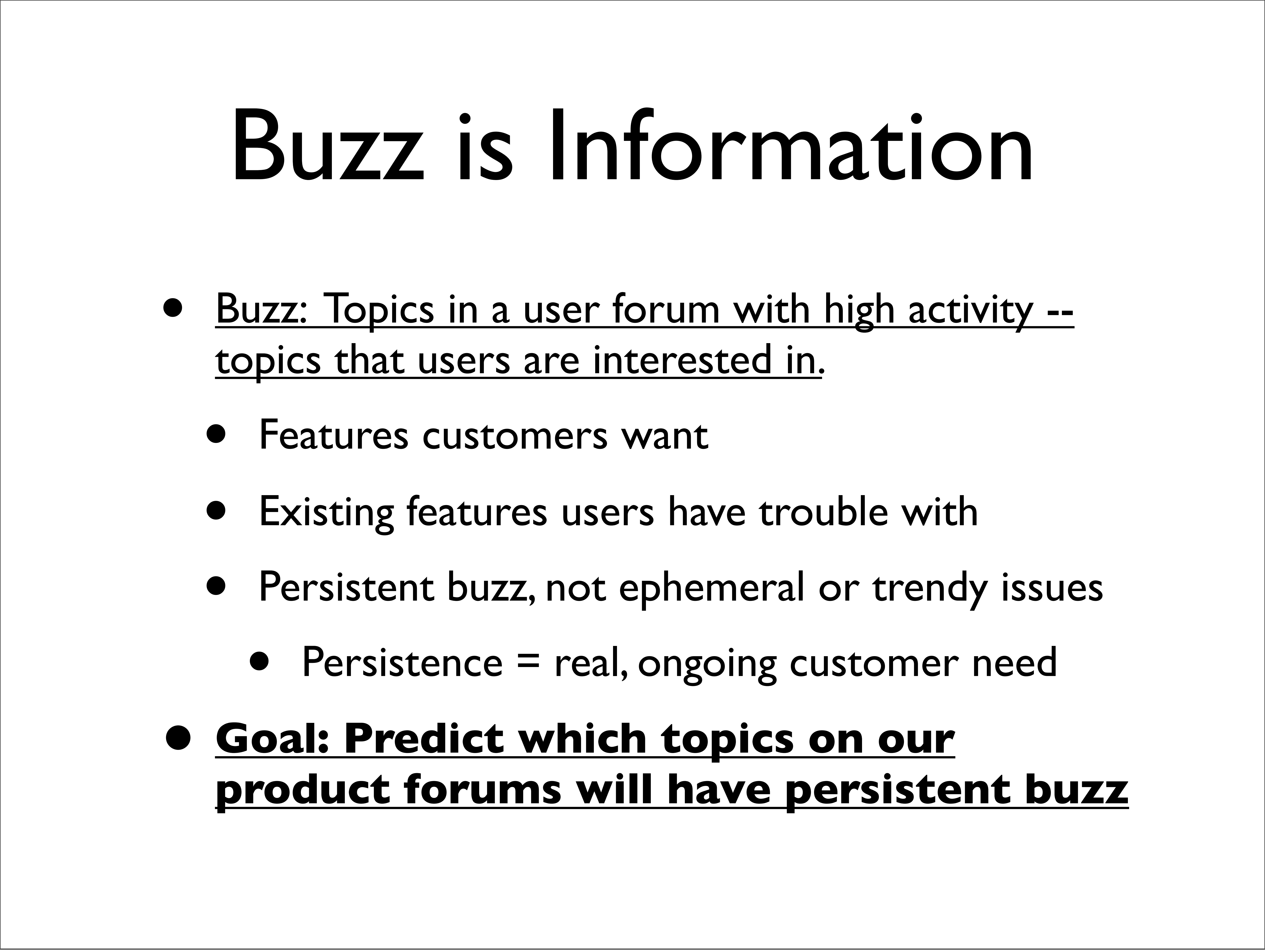
Scenario Description

- WVCorp: the company you (the data scientist) work for
  - WVCorp has user forums and discussion boards for each of their products, where customers can discuss issues and features.
  - “Buzz”: when a topic on the user forum has a very high activity level -- considered an indication of user interest in that topic.
- eRead:WVCorp’s ebook reader product
- TimeWrangler: WVCorp’s time-management app
- BookBits: A competitor’s ebook reader product
- GCal: a third-party cloud-based calendar infrastructure that TimeWrangler can integrate with

Notes to describe the fictional world this presentation takes place in

Predicting Buzz on User Forums

WVCorp Data Science Team
Notional Peer Presentation



Buzz is Information

- Buzz: Topics in a user forum with high activity -- topics that users are interested in.
  - Features customers want
  - Existing features users have trouble with
  - Persistent buzz, not ephemeral or trendy issues
    - Persistence = real, ongoing customer need
- **Goal: Predict which topics on our product forums will have persistent buzz**

Peer audiences are generally more interested in the prediction task, and will accept mild motivation (“buzz is useful”). However, for some audiences (especially to peer groups outside your organization), you may want to lead with the business problem (the first slide or two of the project sponsor presentation) for context.

Related Work

- Predicting movie success through social network and sentiment analysis
    - Krauss, Nann, et.al. *European Conference on Information Systems*, 2008
  - IMDB message boards, Box Office Mojo website
  - Variables: discussion intensity, positivity
- Predicting asset value (stock prices, etc) through Twitter Buzz
    - Zhang, Fuehres, Gloor, *Advances in Collective Intelligence*, 2011
  - Time series analysis on pre-chosen keywords

Academic peer presentations generally have a related work section: discuss others who have done research on related problems, and how your approach is similar to theirs, and how it differs (and perhaps why you didn't take the same approach they did).

Pilot Study

- Collected three weeks of data from forum
  - 7900 topics, 96 variables
    - 791 topics held out for model evaluation
  - 22% of topics in Week 1 of the data set buzzed in Weeks 2/3
- Trained Random Forest on Week 1 to identify which topics will buzz in Weeks 2/3
  - Buzz = Sustained increase of 500+ active discussions in topic/day, relative to Week 1, Day 1
- Feedback from team of five product managers -- how useful were the results?

High level introduction to what we did. Leave in the bullet about the product managers, just to keep “relevance” of model in the discussion

Model Variables

- We started with metrics already monitored by system.
  - #Authors/topic
  - #Discussions/topic
  - #Displays of topic to forum users
  - Average #contributors to a discussion in the topic
  - Average discussion length in a topic
  - How often a discussion in a topic is forwarded to social media

- Obviously problematic -- only point measurements
  - Ideally, we want to measure evolution
    - Are, e.g. the number of authors increasing/decreasing? How fast?
  - Time-series analysis
  - How well can we do with what we have?

For this audience, lots of details.

Random Forest Model

- Efficient on large data, large number of input variables
- Few prior assumptions on variable distribution/ interactions
- We limited complexity to reduce overfit
  - 100 nodes/tree maximum
  - Minimum node size 20
  - More data would eliminate the need for these steps

Details about the modeling approach chosen

Results

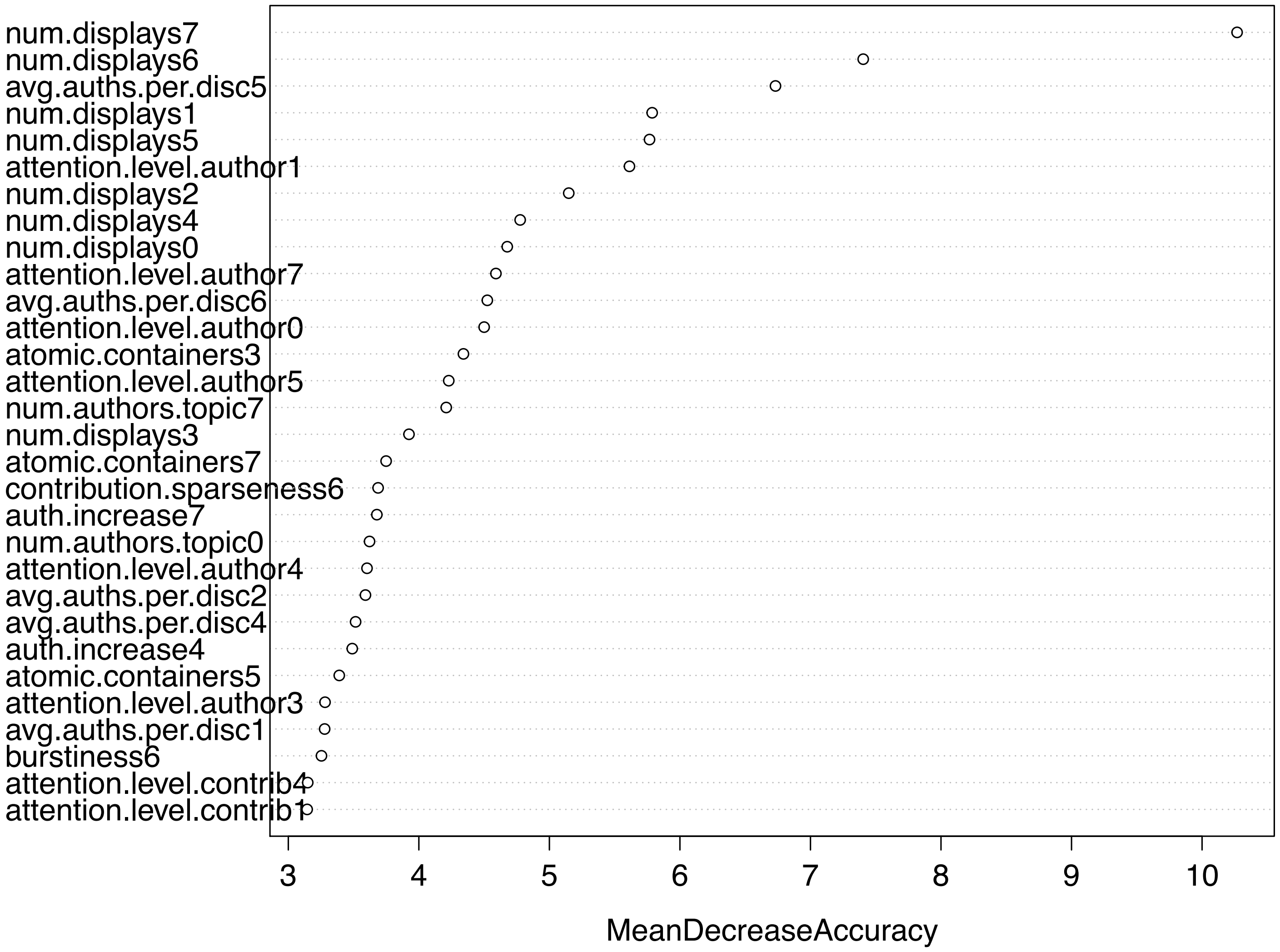
- 84% recall, 83% precision
- Reduced manual scan of forums by over a factor of 4
  - From 791 to 184 topics to inspect
- PMs: 75% of identified topics produced "valuable insight"

	Predicted No Buzz	Predicted Buzz	
No Buzz	579	35	614
Buzz	28	149	177
Total	607	184	791

Model performance, and other related facts

Variable Importance

- Key inputs:
  - # times topic is displayed to user (num.displays)
  - # authors contributing to topic (attention.level.author)
- Velocity variables for these two inputs could improve model



Discussion of input variables and their impact on the model

Example Discovery

- Topic: TimeWrangler →GCal Integration
  - # discussions up since GCal v. 7 release
    - GCal events not consistently showing up; mislabeled.
    - TimeWrangler tasks going to wrong GCalendar
  - **Hot on forums before hot in customer support logs**
    - Forum activity triggered the model two days after GCal update
    - Customer support didn't notice for a week

Example result.

Future Work

- Better input variables
  - Shape and velocity variables
    - How quickly #authors grows/shrinks
    - How much #topic displays increases/decreases
  - Information about new forum visitors
    - What questions do first-time visitors come to ask?
- Research optimal model retraining schedule

Thank You